Expert level evaluations for explainable AI (XAI) methods in the medical domain

Muddamsetty, Satya Mahesh; Jahromi, Mohammad Naser Sabet; Moeslund, Thomas B.

Published in:
Pattern Recognition. ICPR International Workshops and Challenges

DOI (link to publication from Publisher):
10.1007/978-3-030-68796-0_3

Publication date:
2021

Document Version
Accepted author manuscript, peer reviewed version

Expert level evaluations for explainable AI (XAI) methods in the medical domain

Satya M. Muddamsetty[0000−0003−0935−4609], Mohammad N. S. Jahromi[0000−0002−6332−7567], and Thomas B. Moeslund[0000−0001−7584−5209]

Visual Analysis of People Laboratory (VAP), Aalborg University, Rendsburggade 14, 9000 Aalborg, Denmark
{smmu,mosa,tbm}@create.aau.dk

Abstract. The recently emerged field of explainable artificial intelligence (XAI) attempts to shed lights on 'black box' Machine Learning (ML) models in understandable terms for human. As several explanation methods are developed alongside different applications for a black box model, the need for expert-level evaluation in inspecting their effectiveness becomes inevitable. This is significantly important for sensitive domains such as medical applications where evaluation of experts is essential to better understand how accurate the results of complex ML are and debug the models if necessary. The aim of this study is to experimentally show how the expert-level evaluation of XAI methods in a medical application can be utilized and aligned with the actual explanations generated by the clinician. To this end, we collect annotations from expert subjects equipped with an eye-tracker while they classify medical images and devise an approach for comparing the results with those obtained from XAI methods. We demonstrate the effectiveness of our approach in several experiments.

Keywords: Explainable AI (XAI), Deep learning, Expert-level explanation, XAI evaluation, Retinal Images, Eye-tracker.

1 Introduction

Machine Learning (ML) models are becoming an essential part of the current technology due to their ability to outperform humans in solving particular tasks such as spam detection [8, 15], healthcare [5], ophthalmology [24] and autonomous robots [27]. Furthermore, the ML model can be employed to help experts in supporting their decision in domains such as medical or risk analysis where actionable solutions may have serious consequences [13, 26]. Recent advances in ML promise to improve retinal diseases screening substantially and to improve diagnosis accuracy. Systems developed using these methods have demonstrated expert-level accuracy in diagnosis for multiple eye diseases including diabetic retinopathy [20], age related macular degeneration (AMD) [11], glaucoma [18] and other anomalies associated with retinal diseases, and to monitor their progression. However, the impact of these models in clinical settings

is not completely understood. Previous attempts to use ML algorithms in a computer-assisted diagnosis setting have faced numerous challenges, including both over reliance (repeating errors made by the model) and under reliance (ignoring accurate algorithm predictions) [14, 22]. Some of these issues may be avoided if the computer assisted diagnosis system can explain its black box AI predictions [3]. Explainable AI (XAI) aims at decoding the decision of AI (Deep learning/Machine learning) black box to the extent of human-interpretable level. For instance, if we are to use AI algorithms to classify Diabetic Retinopathy (DR) levels from retinal fundus images, can the algorithm generate further interpretable justification for its prediction results? Can that justification be presented visually? Is that visualization aligned closely with expert explanation? Generally, in sensitive domains such as clinical settings, the domain experts (clinicians) are skeptical in supporting interpretations generated by AI diagnosis tools as a result of high involved risk [6, 26]. But, if instead of developing various explanation methods for the sensitive domains, the effectiveness of their evaluation method is studied when expert subjects are involved in loop, then AI diagnosis tool gain further trust by the domain experts. Therefore, in addition to improving accuracy of such a tool, the notion of trust, need for transparency and robustness implies how crucial it is to study the effect of expert evaluation in the context of XAI methods.

XAI evaluation methods are broadly classified into three categories [9]: They are, Application-Grounded Evaluation, Human-Grounded Evaluation and Functionally Grounded Evaluation. Application-Grounded Evaluation quantifies how expert-generated explanation can properly help other humans in specific tasks. The quality of this evaluation is tested by employing domain experts to accomplish certain tasks within the context of an application. For example, an ophthalmologist should evaluate a diagnosis system in determining the DR level from retinal fundus images. On the other hand, in the *Human-Grounded Evaluation* evaluations are done using non-expert humans on simplified tasks. For instance non-experts or users will be shown different explanations and the user would choose the best one. The authors in [4, 23] evaluated their method using non-experts, asking to identify which XAI method provides good explanation. The Functionally-Grounded Evaluation discussed in [8] is basically independent of human subject. Most of the state-of-art methods falls into this category [1, 21]. For example, the authors in [19] proposed casual metrics *insertion* and *deletion* which, are independent of humans to evaluate the faithfulness of the XAI methods. The intuition behind the *deletion* and *insertion* metrics is that the removal or inserting of the 'cause' will make the AI model to change its decision. However, functionally and human grounded evaluations will not be suitable for such sensitive medical domains. In practice, all types of evaluation have equal importance. Choosing a right evaluation method is subject to the explanation context. For instance, if we seek to generate an explanation that is limited to experts or specific application such as a medical diagnosis tool, the Application-grounded evaluation could become more appropriate. This is due to fact that

for such a unique application careful expert studies are required. Therefore, to address the specific medical case such as screening the retinal diseases across retinal fundus images, it is necessary to involve domain experts within this field to evaluate the explanations of black box models predictions. One way of performing this task is through an interactive collection of the expert feedback on generating actual explanation using an eye-tracker.

To this end, the main contribution of this paper is to develop a collection of eye-tracking data from 3 expert subjects across 150 retinal fundus images for medical application. Concretely, domain clinicians are equipped with an eye-tracker in an interactive experimental settings to understand how they classify retinal diseases and assess the retinal image quality. The experts evaluate the five DR level (No DR, Mild, Moderate, Severe and Proliferative DR) and the retinal image quality (Good/Bad), respectively. Finally the heatmaps obtained via the eye-tracker can be compared directly by XAI methods. In this work we use heatmaps generated together by two XAI state-of-the-art methods namely SIDU [17], GRAD-CAM [23] using two different evaluation metrics.

The rest of the paper is organized as follows. In Section 2 we describe the eye-tracking experiments and data collection. Section 3 describes the XAI methods used for evaluation. Section 4 describes comparison metrics for XAI methods and Section 5 shows performance evaluation of XAI methods and comparisons. Finally, Section 6 provides concluding remarks.

2 Eye-tracking Experiments and Data collection

In this section, we discuss how we employ expert subjects to collect annotation of medical images with an eye-tracker in an interactive experiment setting. The experts annotation experiments consist of two phases. In the first phase a total of 100 images are randomly drawn from the Retinal Fundus Image Quality Assessment (RFIQA) dataset [16]. Each expert subject (ophthalmologist) is then required to classify each image (stimuli) as a Good/Bad quality. During each session of the experiment an ophthalmologist is equipped with an eye-tracker. In a similar setting, the second phase is conducted to highlight salient regions corresponding to Diabetic Retinopathy (DR) levels (5 grades) in the eye fundus Images. In this setup, in order to utilize different dataset for evaluation, 50 images are selected randomly from the EyePacs dataset [10]. To ensure variance in our experiment, we asked ophthalmologist experts from medical community to participate in the experiments. The data collection protocol as well as hardware setup are discussed in the subsequent section. Note that the collection procedures are identical in both phases.

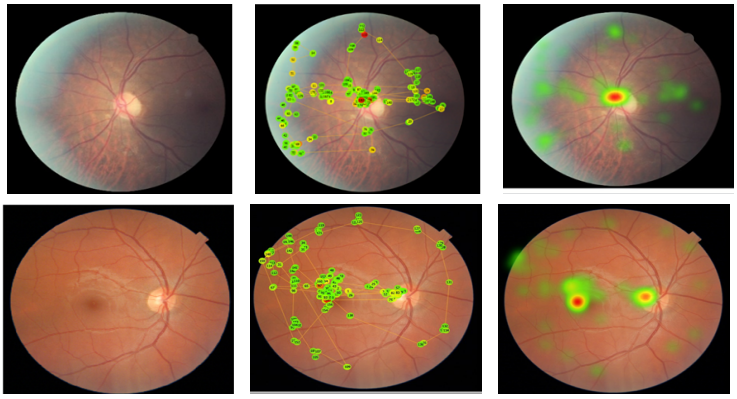
2.1 Data collection protocol

In order to record the eye-fixation (spatial coordinate on the screen) of each expert on the fundus images, we utilized Tobii-X120 eye-tracker [25] as follows:

1. Instructions were given to the expert to sit in front of a screen where the eye-tracker was attached to the base stand of a monitor and facing a subject.
2. In order for the eye-tracker to capture the eye-fixation properly, the subject has to sit within 60 cm distance from the screen. This distance was measured in the IMOTION (Software) [7] where the ultimate data annotation took place.
3. After a careful calibration of the eye-tracker setup done in the IMOTION, a block of stimuli displayed on the screen for the expert to evaluate the good/bad quality of each stimuli (Natural image). Each block of data composed of 25 cells and in each cell, three images were located. The cross-fixation (+) image at in the beginning (1 sec duration) in order to refresh the visionary system between each stimuli transition. The second image was the main stimuli presented to the expert and finally a survey with the task question (Good/Bad quality or DR level assessment).
4. Once the eye-fixation were collected for the first block (25 stimuli), a break is given for 10-minutes. This process were repeated for all the stimuli.

Figure 1 illustrates the eye-tracking collection setup with the eye-tracker positioned below the screen and facing to the experts during the experiment. In addition, Figure 2 shows samples of recorded eye-fixation and their generated heatmaps for an expert subject. Note that the heatmaps generation are done via fitting a Gaussian kernels at each coordinate location of eye-fixation (third column in the figure).

Fig. 1. Eye-tracking data collection from experts for screening the diabetic retinopathy and retinal images quality levels.



(a) Original Image (b) Eyefixations (c) Heatmaps

Fig. 2. Eye-tracking data samples of recorded eye-fixation and their generated heatmaps collected from experts for screening the diabetic retinopathy and retinal images quality levels.

3 Explainable AI Methods

To evaluate the visual explanation on the proposed dataset we considered two most recent visual explanation methods. They are briefly described in sections 3.1 and 3.2.

3.1 SIDU

A new visual explanation method known as SIDU proposed recently in [17] estimates the pixel saliency by extracting the last convolutional layer of the deep CNN model and creating the similarity difference mask which is eventually combined to form a final map for generating the visual explanation of the prediction. This method generates a heatmap based on two steps: Similarity difference and Uniqueness. First, a heatmap of the most salient areas of an image is generated by calculating the similarity difference between sets of feature activation maps. Secondly, it evaluates feature map uniqueness. This step calculates how different a specific feature map is from the others. If a feature map is unique, then it will be labelled as more salient and have a higher weight. The final score that gives the feature importance is given by the dot product between the two values, which is then used to calculate the weighted sum of all feature activation image masks and generate the visual explanation. It was shown via the quantitative and qualitative (human trust) experiments that for both general and critical medical data, the SIDU method outperforms state-of-the-art [17]. The ability of properly localizing the region of interest in the clinical eye fundus images makes SIDU a well-suited method to provide transparent explanation and audit model output that is crucial for sensitive domains such as medical diagnosis.

3.2 GRAD-CAM

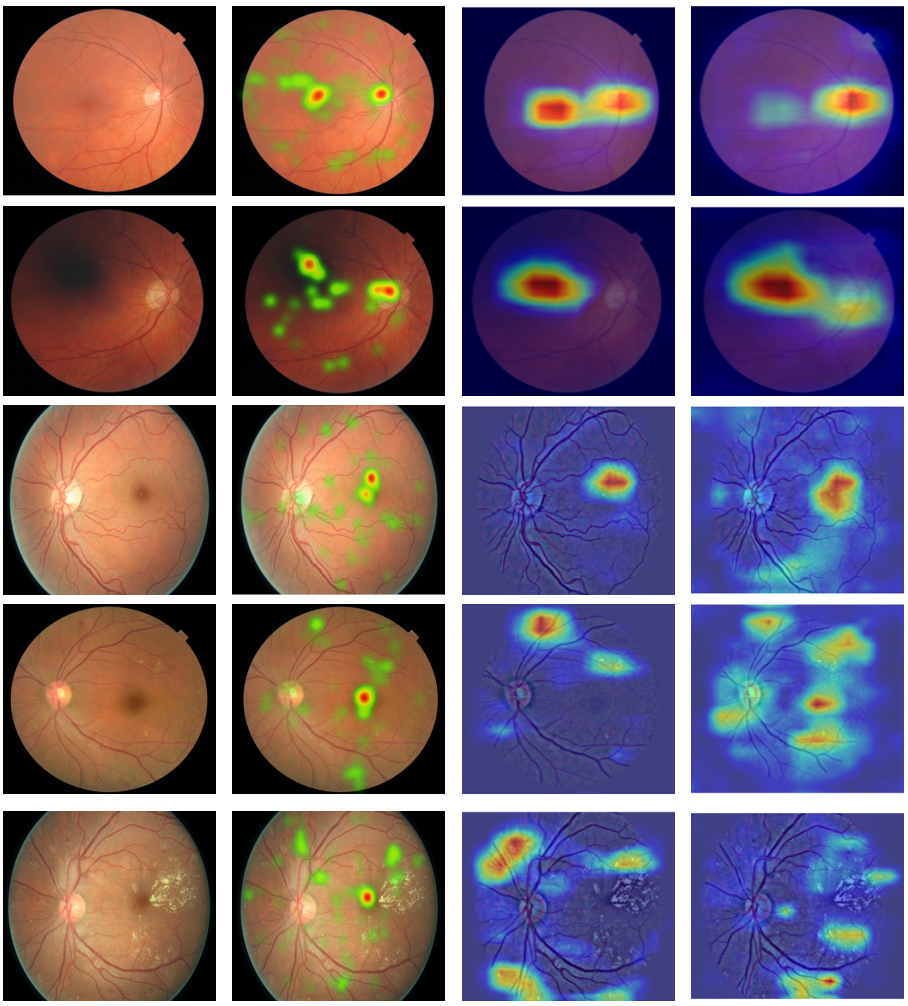
Grad-CAM is a method which generates visual explanations via gradient based localization [23]. It extracts the gradients from the last convolution layer of the network. The intuition behind this method is that the layer prior to the classification retains the information of feature relevance while maintaining spatial relations, and therefore it can generate a heatmap (based on a weighted combination of activation maps dependent on gradient score), which highlights the features with a positive influence for the specific class that is chosen as the prediction. Given any CNN model, Grad-CAM is an class-discriminative localization technique which can generate visual explanations without requiring architectural changes or re-training.

4 Comparison Metrics for XAI methods

To compare heatmaps we choose the two usual metrics used in state-of-the-art methods for evaluation of saliency detection [2]. The main reason for choosing more than one evaluation measure is to ensure that the discussion about the results is as independent as possible from the choice of the metrics. The results of the different evaluation metrics are not necessarily the same, but when two metrics show similarities, then it is easy to interpret the robustness of the methods. These metrics are used do evaluate the performance of the XAI methods are two different kinds of experiments described in section 5.2.

4.1 Area under ROC Curve (AUC)

The Receiver Operating Characteristics (ROC) measure is one of the most popular and most widely used method in the community for assessing the degree of similarity of two saliency maps and it measures the trade-off between true and false positives at different discrimination threshold values (level sets) [2]. It is a graphical plot which describes the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the total actual positives (TPR = true positive rate) versus the fraction of false positives out of the total actual negatives (FPR = false positive rate), at various threshold values. A good prediction method would give a TPR of 1 at a FPR of 0, yielding a point in the upper left corner of the ROC space that corresponds to a perfect classification. A completely random guess would give a point along a diagonal line from the left bottom to the top right corner. The diagonal divides the ROC space and points above the diagonal represent good classification results (better than random), points below the line poor results (worse than random). Thus, a measure of performance derived from the ROC curve is the AUC (Area Under Curve) which is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative'). The XAI visual explanation heatmap is treated as a binary classifier of fixations at various threshold values (level sets), and an ROC curve is swept out by measuring the true and false positive rates under each binary classifier (level set).



(a)Original Image (b) Eye-tracker (c) GRAD-CAM [23] (d) SIDU [17]

Fig. 3. Comparison of XAI methods visual explanation with human visual explanation (heatmaps). First two rows describes the explanations of good and bad quality of retinal images. Third, fourth and fifth row describes the explanations of mild, moderate and sever DR levels. In a real scenario, the ophthalmologists inspect the image quality or DR levels by looking around exact regions (areas of the heatmaps captured by the eye-tracker in the 2^{nd} column) of the eye fundus images. The generated heatmaps in 3^{rd} and 4^{th} columns by the GRAD-CAM and SIDU demonstrate how the visual explanation methods are closely aligns with human experts.

4.2 Kullback-Leibler Divergence (KL-DIV)

The Kullback-Leibler Divergence is a metric, which estimates dissimilarity between two probability density functions [2]. To evaluate the XAI methods, the distributions are given by the eye-fixations points and the heatmap (visual explanation maps) produced by the model. Let FM be the probability distribution of the heat map from eye tracking data, and EM be the probability distribution of the visual explanation map. The distributions are normalized and they are given by :

$$EM(x) = \frac{EM(x)}{\sum_{x=1}^{X} EM(x) + \epsilon}, \tag{1}$$

$$FM(x) = \frac{FM(x)}{\sum_{x=1}^{X} FM(x) + \epsilon}, \tag{2}$$

where X is the number of pixels and ϵ is a regularization constant to avoid division by zero. The KL-divergence measure is non-linear and varies in the range of zero to infinity. The lower score indicates that the EM maps have better approximation of the human expert eye-fixation ground truth.

5 Experimental Evaluation and Results

In this section we perform the experimental evaluation of two state-of-the-art XAI methods described in section 3 on proposed datset. We conducted two experiments, for the given trained CNN models. In first experiment, we evaluated the explanations of retinal fundus image quality predictions and in the second experiment we evaluated the explanations of Diabetic Retinopathy disease level classification.

5.1 Training CNN Models

To evaluate the XAI methods, we trained two CNN models on two different datasets. First, we trained the existing ResNet50 [12] with an additional two FC layers and softmax layer on the Retinal Fundus Image Quality Assessment (RFIQA) dataset from the medical domain. The dataset consists of 9,945 images with two levels of quality, 'Good' and 'Bad'. The retinal images were collected from a large number of patients with retinal diseases [16]. The dataset is split into 80% training, 10% validation and 10% testing. Data augmentation is performed on the training samples. We apply image transformations such as random rotation, width shift, height shift, zooming, horizontal flipping and scaling to the RFIQA training subset to enlarge the dataset. The CNN model is trained over 120 epochs with batch size of 3. We use categorical cross entropy as a loss function and SGD as optimizer with learning rate 10^{-4} and momentum as 0.9. The CNN model is initialized with imagenet weights and trained the whole model on

the RFIQA dataset. The CNN model achieves 94% accuaracy. The explanation methods uses the trained model for explaining the prediction of the RFIQA test subset with 1028 images whereas the second CNN model is trained for classifying the diabetic retinopathy (DR) disease levels. We used existing ResNet50 [12] model similar to first one. The CNN model is trained on the EyePacs dataset [10]. we initialized with ImageNet pre-trained weights and trained the whole model parameters on EyePacs dataset. The CNN model achieves 85% on test dataset which 10k images of five levels of DR. Both model have 'human like' performance and hence XAI generated heatmaps can be expected to behave similar to those of human expert. The frameworks are implemented on Tensorflow keras with GPU memory of $11GB$, Nvidia, RTX 2080Ti.

5.2 Results and Discussion

In the first experiment, we use the RFIQA images eye-tracking data recordings described in section 2 to generate and evaluate the explanation by the XAI algorithms. To this end, we first generate ground truth heatmaps by applying Gaussian distributions on human expert eye-fixations. These heatmaps are then used to compare with the XAI heatmaps. Table 1 summarizes the results obtained by two different XAI methods on our proposed RFIQA eye-tracking data using AUC and KL-DIV evaluation measures as described in section 4.1 and 4.2 respectively. For the AUC measure, we observe that, GRAD-CAM [23] shows slightly better performances compared to SIDU [17] for both the experts, whereas for the KL-DIV measure both methods performed equally well for expert 1 and for expert 2 GRAD-CAM [23] has shown better performance.

In the second experiment we evaluated XAI methods using DR disease levels eye-tracking data recording described in section 2. We follow a similar procedure for generating the ground truth heatmaps. We collected the eye-fixations from three experts and compared the XAI methods, explanations with the three experts individually. Table 2 summarizes the results obtained by two different XAI methods on our proposed eye-tracking data using two evaluation measures. We observe that SIDU [17] performs better than GRAD-CAM [23] for both AUC and KL-DIV measures for all the experts. Figure 3 shows visual explantions comparisons of XAI methods with human visual explanation of Good, Bad, quality grades and DR disease levels such as No DR, Mild, Moderate, Severe, Proliferative DR levels predictions. From the figure, we can clearly observe that the visual explanations from the XAI methods are closely aligns with human experts. In practice, the doctors verify the visibility of the optical disc and macular regions in a good quality image, corresponding to the highlighted regions in the heatmap 1^{st} row. Similarly, the bad quality image, 2^{nd} row is due to the shadow just above the center of the image, i.e., exactly the region highlighted by the XAI methods. For the DR levels the practitioners look in the lesions such as Microaneurysms (tiny red lesions), Haemorrhages (Bright red lesions), Exudates (Yellow spots) near the optical disk, the macula and the region surrounding to macula and this can be observed in the heatmaps of human experts and XAI

methods 3^{rd}, $4^{th}row$ and $5^{th}row$ with mild, moderate and Severe DR levels. Therefore from the above two experiments we can conclude that the evaluation of XAI methods in medical domain requires human experts for gaining greater trust and transparency.

Table 1. Evaluation of XAI methods on proposed eye-tracking dataset using AUC and KL-DIV on retinal images quality levels dataset.

XAI methods	AUC measure ↑			KL-DIV measure↓		
	Expert1	Expert 2	Expert3	Expert1	Expert2	Expert3
SIDU [17]	0.6545	0.5899	0.6442	11.7712	14.1240	12.2316
GRAD-CAM [23]	**0.6605**	**0.6125**	**0.6575**	**11.6087**	**13.3929**	**11.7866**

Table 2. Evaluation of XAI methods on proposed eye-tracking using AUC and KL-DIV on DR disease levels dataset.

XAI methods	AUC measure ↑			KL-DIV measure↓		
	Expert1	Expert 2	Expert3	Expert1	Expert2	Expert3
SIDU [17]	**0.6089**	**0.5805**	**0.5834**	**12.9420**	**14.14708**	**13.6627**
GRAD-CAM [23]	0.5734	0.5454	0.5504	14.0974	15.2675	14.6741

6 Concluding Remarks

In this paper we proposed a framework for evaluating explainable AI (XAI) methods using an eye-tracker in the medical domain particularly for the screening of retinal diseases, DR and quality assessment for retinal images. It is designed specifically for evaluating XAI methods in the medical domain. To the best of our knowledge, the proposed eye-tracker dataset is the first of its kind for evaluating the visual explanations in the medical domain by involving human experts (ophthalmologists). Experimental results using two different datasets with different characteristics show the importance of involving human experts in evaluating XAI methods.

References

1. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one **10**(7) (2015)
2. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? IEEE transactions on pattern analysis and machine intelligence **41**(3), 740–757 (2018)
3. Cabitza, F., Rasoini, R., Gensini, G.F.: Unintended consequences of machine learning in medicine. Jama **318**(6), 517–518 (2017)
4. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 839–847. IEEE (2018)
5. Chen, R., Yang, L., Goodison, S., Sun, Y.: Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data. Bioinformatics **36**(5), 1476–1483 (2020)
6. Chromik, M., Schuessler, M.: A taxonomy for human subject evaluation of black-box explanations in xai. In: ExSS-ATEC@ IUI (2020)
7. De Lemos, J.: Visual attention and emotional response detection and display system (Nov 15 2007), uS Patent App. 11/685,552
8. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
9. Doshi-Velez, F., Kim, B.: Considerations for evaluation and generalization in interpretable machine learning. In: Explainable and Interpretable Models in Computer Vision and Machine Learning, pp. 3–17. Springer (2018)
10. EyePACS: Diabetic retinopathy detection of kaggle. Available in: https://www.kaggle. com/c/diabetic-retinopathy-detection/data (2015)
11. Grassmann, F., Mengelkamp, J., Brandl, C., Harsch, S., Zimmermann, M.E., Linkohr, B., Peters, A., Heid, I.M., Palm, C., Weber, B.H.: A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. Ophthalmology **125**(9), 1410–1420 (2018)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Hengstler, M., Enkel, E., Duelli, S.: Applied artificial intelligence and trust—the case of autonomous vehicles and medical assistance devices. Technological Forecasting and Social Change **105**, 105–120 (2016)
14. Kohli, A., Jha, S.: Why cad failed in mammography. Journal of the American College of Radiology **15**(3), 535–537 (2018)
15. Lin, C.F.: Application-grounded evaluation of predictive model explanation methods
16. Muddamsetty, S.M., Moeslund, T.B.: Multi-level quality assessment of retinal fundus images using deep convolutional neural network. Submitted to VISAPP (2021)
17. Muddamsetty, S.M., Mohammad, N.S.J., Moeslund, T.B.: Sidu: Similarity difference and uniqueness method for explainable ai. In: 2020 IEEE International Conference on Image Processing (ICIP). pp. 3269–3273 (2020). https://doi.org/10.1109/ICIP40778.2020.9190952

18. Nayak, J., Acharya, R., Bhat, P.S., S., N., Lim, T.: Automated diagnosis of glaucoma using digital fundus images. Journal of medical systems **33**(5), 337 (2009)
19. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. In: Proceedings of the British Machine Vision Conference (BMVC) (2018)
20. Raman, R., Srinivasan, S., Virmani, S., Sivaprasad, S., Rao, C., Rajalakshmi, R.: Fundus photograph-based deep learning algorithms in detecting diabetic retinopathy. Eye (11 2018). https://doi.org/10.1038/s41433-018-0269-y
21. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.R.: Evaluating the visualization of what a deep neural network has learned. IEEE transactions on neural networks and learning systems **28**(11), 2660–2673 (2016)
22. Sayres, R., Taly, A., Rahimy, E., Blumer, K., Coz, D., Hammel, N., Krause, J., Narayanaswamy, A., Rastegar, Z., Wu, D., et al.: Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. Ophthalmology **126**(4), 552–564 (2019)
23. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
24. Son, J., Shin, J.Y., Kim, H.D., Jung, K.H., Park, K.H., Park, S.J.: Development and validation of deep learning models for screening multiple abnormal findings in retinal fundus images. Ophthalmology **127**(1), 85–94 (2020)
25. Technology, T.: User manual: Tobii x60 and x120 eye trackers (2008)
26. Weld, D.S., Bansal, G.: The challenge of crafting intelligible intelligence. Communications of the ACM **62**(6), 70–79 (2019)
27. You, C., Lu, J., Filev, D., Tsiotras, P.: Advanced planning for autonomous vehicles using reinforcement learning and deep inverse reinforcement learning. Robotics and Autonomous Systems **114**, 1–18 (2019)